

EXHIBIT B

Page 1

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>



MICROSOFT / TECH / ARTIFICIAL INTELLIGENCE

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer

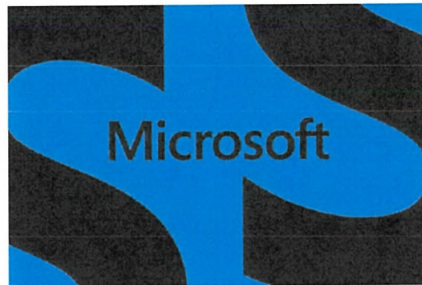


Illustration: The Verge

/ Microsoft says it connected tens of thousands of Nvidia A100 chips and reworked server racks to build the hardware behind ChatGPT and its own Bing AI bot.

By EMMA ROTH
 Mar 13, 2023, 2:03 PM EDT | 18 Comments | 15 New



If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)

Microsoft spent hundreds of millions of dollars building a massive supercomputer to help power OpenAI's ChatGPT chatbot, according to a report from *Bloomberg*. In a pair of blog posts published on Monday, Microsoft explains how it created Azure's powerful artificial intelligence infrastructure used by OpenAI and how its systems are getting even more robust.

To build the supercomputer that powers OpenAI's projects, Microsoft says it linked together thousands of Nvidia graphics processing units (GPUs) on its Azure cloud computing platform. In turn, this allowed OpenAI to train increasingly powerful models and "unlocked the AI capabilities" of tools like ChatGPT and Bing.

Scott Guthrie, Microsoft's vice president of AI and cloud, said the company spent several hundreds of millions of dollars on the project, according to a statement given to *Bloomberg*. And while that may seem like a drop in the bucket for Microsoft, which recently extended its multiyear, multibillion dollar investment in OpenAI, it certainly demonstrates that it's willing to throw even more money at the AI space.

Microsoft's already working to make Azure's AI capabilities even more powerful

Microsoft's already working to make Azure's AI capabilities even more powerful with the launch of its new virtual machines that use Nvidia's H100 and A100 Tensor Core GPUs, as well as Quantum-2 InfiniBand networking, a project both companies teased last year.

According to Microsoft, this should allow OpenAI and other companies that rely on Azure to train larger and more complex AI models.

"We saw that we would need to build special purpose clusters focusing on enabling large training workloads and OpenAI was one of the early proof points for that," Eric Boyd, Microsoft's corporate vice president of Azure AI, says in a statement. "We worked closely with them to learn what are the key things they were looking for as they built out their training environments and what were the key things they need."

16 COMMENTS (16 NEW)

- 1 If you're diabetic, don't wait for your smartwatch to replace your needles
- 2 Microsoft's new Copilot will change Office documents forever
- 3 Amazon's Swarm is so close to being brilliant
- 4 Amazon's latest layoffs cut 9,000 more jobs in divisions including Twitch and AWS
- 5 This Apple Pencil clone provides 80 percent of the experience for a quarter of the price

Verge Deals / Sign up for Verge Deals to get deals on products we've tested sent to your inbox daily.

Enter your email

SIGN UP

By submitting your email, you agree to our [privacy policy](#) and [terms of service](#). This site is protected by reCAPTCHA and the Google [Privacy Policy](#) and [Terms of Service](#) apply.


Page 2
 Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>

More from this stream [Bing, Bard, and ChatGPT: AI chatbots are resitting the internet](#)


- **Apple has generative AI plans, too.**
Mar 15, 2023, 11:17 AM EDT
- **Microsoft's AI shortcut is reaching more Windows taskbars.**
Mar 15, 2023, 11:10 AM EDT
- **What's new with GPT-4 – from processing pictures to acing tests**
Mar 15, 2023, 10:42 AM EDT
- **Google-backed Anthropic launches Claude, an AI chatbot that's easier to talk to**
Mar 14, 2023, 09:10 PM EDT

[SEE ALL THE STORIES](#)


SPONSORED CONTENT




Google Chrome Users Can Now Block All Ads (Do it Now For Free!)
Paul Tach Type
[Read more](#)




This Is The Most Realistic Game In 2023
Paul Shadon Legends
[Play game](#)




New Hampshire Say Bye To Your Home Insurance Bill If You Live In These
your state/your home insurance
[Learn More](#)



Incredibly: Most Chrome Users Didn't Know How To Block Ads Instantly
Chrome Hacks Tips
[Learn More](#)



Wall St Legend: Once In A Generation Day Is Coming, Prepare Now
Financially Profit
[Learn More](#)



Cardiologist: Too Much Belly Fat? Do This Before Bed
Healthtips/News/Press
[Learn More](#)

Page 3

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>

Today's Storystream
FEED REFINED 35 MINUTES AGO
 HAVE YOU CHANGED YOUR TWITTER 2FA?

Nissan Ariya first drive: an EV pioneer regains its credibility
ROBERTO BALDWIN 72 MINUTES AGO

How Apple's avoiding layoffs: delaying bonus payments, pausing hiring, cutting travel budgets. Sure, Apple didn't go on the "hiring binges" during the pandemic that some of its rivals did. But it also understands something its rivals don't: layoffs are a strong signal that the c-suite screwed up.
Apple Cost-Cutting Efforts: No Layoffs, But Less Travel and Delayed Bonuses
 [BLOOMBERG.COM]

What gadget should we X-ray? ICYMI, we're now sticking our gadgets into a CT scanner that spits out ghostly 3D images of their insides. We've got more videos like this Polaroid on the way — but if there's some small gadget you think *The Verge's* audience would love to see scanned, hit me up! I'm at sean@theverge.com.

Amazon's layoffs included 'just over 400' job cuts at Twitch. Twitch CEO Dan Clancy, who just stepped into the role on Thursday, shared the number as part of a new blog post shortly after the announcement of Amazon's next huge wave of layoffs.
"Like many companies, our business has been impacted by the current macroeconomic environment, and user and revenue growth has not kept pace with our expectations," according to Clancy. "In order to run our business sustainably, we've made the very difficult decision to shrink the size of our workforce."

MOST POPULAR

- If you're diabetic, don't wait for your smartwatch to replace your needles
VICTORIA SONG MAR 15
- Microsoft's new Copilot will change Office documents forever
TOM WARREN MAR 17
- Amazon's Swarm is so close to being brilliant
CHARLES FULHAM-MOORE MAR 17
- Amazon's latest layoffs cut 9,000 more jobs in divisions including Twitch and AWS
EMMA ROTH MAR 20
- This Apple Pencil clone provides 80 percent of the experience for a quarter of the price
DAN SEIFERT MAR 18

decision to shrink the size of our workforce." You can read more about Amazon's layoffs in our post from earlier today.

Amazon's latest layoffs cut 9,000 more jobs in divisions including Twitch and AWS
EMMA ROTH 11:00 AM EDT

YouTube Music contractors will start union vote on Wednesday. Alphabet has appealed the National Labor Relations Board's decision that it counts as a joint employer for the workers, but the election is still happening. It's set to start on the 22nd, according to the Alphabet Workers Union-CWA.
You can catch up on the story so far here:

Contractors who work on YouTube Music are striking
MITCHELL CLARK FEB 9

How are Masayoshi Son's margin loans doing? About a third of Son's shares of Softbank are collateral for margin loans, which are being used to invest in Softbank Vision Fund. The Vision Fund invests in startups, exactly the kind that are vulnerable to the Silicon Valley Bank meltdown.
What happens if there's a margin call on those loans? Great question, can't wait to find out.
 Checking in on SoftBank
 [FINANCIALTIMES]

Page 4

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>


ANDREW WEBSTER TWO HOURS AGO

Greetings from GDC. After a few years away, we're back in person at the Game Developers Conference in San Francisco. All week myself and Verge games reporter Ash Furrish will be sitting in on talks, chatting with developers, and playing upcoming games and bringing the most interesting stories right here. Stay tuned!

TECH

Someone please buy me this glass mouse pad

MONICA CHIN MAR 17



- Microsoft's new Copilot will change Office documents forever
TOM WARREN MAR 17
- Best printer 2023: just buy this Brother laser printer everyone has, it's fine
NILEY PATEL MAR 15

12/16/2024 15:04:31

● Microsoft's AI shortcut is reaching more Windows taskbars.



Paradox reveals Sims competitor Life by You

ANDREW WEBSTER TWO HOURS AGO



Star Wars Jedi: Survivor's new story trailer shows fighting, friends, and the Force

JAY PETERS TWO HOURS AGO



The FTC is apparently closing in on Amazon. We already knew the FTC was investigating Amazon, but *Politico* has laid out just how wide-ranging (and potentially close to fruition) the agency's efforts are. There's also some info on a case it didn't bring: an attempt to block Amazon from acquiring *One Medical* earlier this year.

The FTC extensively investigated the One Medical deal and found evidence of anticompetitive behavior that many at the agency considered damning, but ultimately cleared the deal because they saw it as too hard of a case to win, according to multiple people with knowledge of the agency's thinking.

As *Politico* points out, we're nearing the end of President Joe Biden's first term, however — so if the FTC is going to take action, it may need to happen soon.

Washington prepares for war with Amazon
(POLITICO)

JAY PETERS 11:50 AM EDT

New iPhone 15 Pro renders show that rumored mute button. The fancier iPhones will apparently switch from a mute switch to a mute *button* this year, and new renders in a video spotted by leaker *ShrimpApplePro* reveal that it, well, looks like a button. If this is the real deal, I'm curious if it will be an improvement over the switch, which has been an iPhone staple since the very first one.

In the video, you can also see the iPhone 15 Pro's rumored unified volume button and that the iPhone 15 will apparently keep the mute switch and separate volume buttons.

Page 5
 Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>

More cads images
 The buttons!
 Source in video [pic.twitter.com/xyy/GANCre](https://twitter.com/xyy/GANCre)
 — ChimindudePro (@Nik-hong-lam) March 20, 2023
 — ShrimpApplePro (@VNechoTaco) March 20, 2023

The Google Pixel 7 is on sale for a new low price of \$449
 ANTONIO G. DI BENEDETTO 11:27 AM EDT

United Nations warns 2050 net-zero climate goals are already outdated
 JUSTINE CALMA 11:23 AM EDT

Swarm's co-creator sees the show as an antiheroic parable. Amazon's *Swarm* is about a popstar-obsessed serial killer, but co-creator Janine Nabers also sees the series both as part of the great American antihero canon, and as a way of pushing viewers to ask themselves "Why do white guys get to do all of this?" Why do white guys get to have all the fun, in terms of the Tony Sopranos and the *Breaking Bad* people? "Why indeed."
"Swarm" Boss Janine Nabers Peels Back the Layers of True Crime-Inspired Series and Its Antihero
 [THE HOLLYWOOD REPORTER]

The tech industry's moment of reckoning: layoffs and hiring freezes
 MITCHELL CLARK 11:19 AM EDT

Amazon's latest layoffs cut 9,000 more jobs in divisions including Twitch and AWS
 EMMA BOTH 11:16 AM EDT

Actors you recognize are joining season 2 of Amazon's *Lord of the Rings* show. The first season of *The Rings of Power* was very sumptuous and expensive TV, but most of the actors weren't especially familiar to audiences. Season 2, however, is adding new cast members you might recognize, including Gaius Hinds (who played Dumbledore's brother and Mance Rayder on *Game of Thrones*). As with previous press releases there's no word on who Hinds is playing. So commence with the speculation.
*Gaius Hinds, Rory Kinnear, and Tanya Moodie Join Cast of Prime Video's *The Lord of the Rings: The Rings of Power* for Season Two*
 [PRESS.AMAZONSTUDIOS.COM]

The Internet Archive is defending its digital library in court today
 ADI ROBERTSON 10:59 AM EDT

The *Last of Us* won't be back for a while. We have a bit of a wait on our hands for the second season of HBO's hit show. In a recent interview on *The Jonathan Ross Show*, Bella Ramsey predicted that it isn't likely to arrive until the end of 2024, or early 2025. Episodes are still being written, and there's no firm date when filming will start. "It will be a while. I think we'll probably shoot at the end of this year, beginning of next," she said.
*Bella Ramsey shares grueling update on *Last of Us* season 2*
 [THE INDEPENDENT]

Text-to-video AI inches closer as startup Runway announces new model
 JAMES VINCENT 10:58 AM EDT

Netflix is adding Monument Valley next year as part of its continued gaming push
 ANDREW WRINER 10:50 AM EDT

Today's Vergecast was not created by GPT-4.
 DAVID PIERCE MAR 17

'The Goliath is Amazon': after 100 years, Barnes & Noble wants to go back to its indie roots
 NILAY PATEL MAR 16

Today on The Vergecast: Moon photos, Silicon Valley Bank, Moon photos, ChatGPT, and Moon photos.
 DAVID PIERCE MAR 15

Why Spotify wants to look like TikTok, with co-president Gustav Söderström
 ALEX HEATH MAR 14

Building tiny keycaps into a small business
 ALEX GRANZ MAR 13

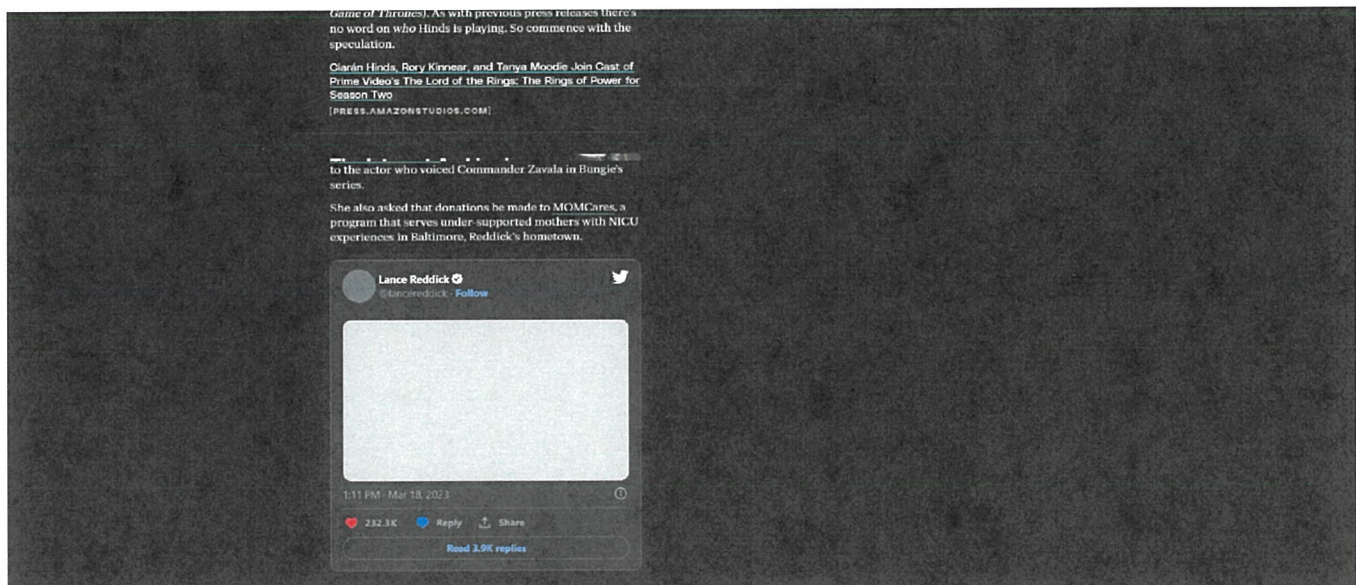
I'm sick of reviewing gorgeous Chromebooks with bad battery
 MONICA CHOI MAR 12

This Apple Pencil sleeve provides 80 percent of the experience for a quarter of the price
 TONY COLETTI MAR 11

The best workout keyboards you can buy right now
 JAMES VINCENT MAR 10

Page 6

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>



Page 7

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>

How Tumblr turned social media polls into a game design challenge

JESS WEATHERBED 9:00 AM EDT

Nordic citizens are again 'happiest' as US climbs to 15th on global list

JESS WEATHERBED 8:50 AM EDT

Microsoft's new Xbox mobile gaming store may launch in 2024

TOM WARREN 8:07 AM EDT

Netflix's ad-supported tier is reportedly gathering momentum in the US

JOE PORTER 3:43 AM EDT

Today's the last day to switch away from Twitter's SMS 2FA method

EMMA ROTH MAR 19

Here's even more evidence Valve is working on a CS:GO update. In case you weren't convinced by previous reports that indicate a CS:GO update is imminent, now PCGamesN has spotted that Valve filed for a "CS2" trademark on March 14th.

This tracks with the "cs2.exe" file names users uncovered in a recent Nvidia driver update. We just don't know whether the game is actually a sequel or an upgraded version using Valve's Source 2 engine.

Counter-Strike 2 trademarks filed by Valve (PCGAMESN)

On The Vergecast: The robot internet is coming. According to Microsoft and Google, this is the future: An AI writes your emails, which are read and summarized by another AI, and then another one synthesizes it and chimes in on the thread, which is then summarized by your AI, and round and round we go.

This week on The Vergecast, it's AI all the way down!

GPT-4 is coming for your work tools — ...

Google Pixel exploit reverses edited parts of screenshots

EMMA ROTH MAR 19

Why Tesla's Full Self Driving... isn't. This story starts with Elon Musk's decision to take radar sensors off Tesla cars, but is ultimately about what happens when an inconsistent, hard-driving management style runs into unexpectedly hard problems. And then what happens when a bunch of your best people disappear to try and fix a flailing social network. It's a good read!

How Elon Musk knocked Tesla's 'Full Self-Driving' off course (WASHINGTON POST)

How many A-button presses does it take to beat Super Mario 64? Just 13, apparently.

To see how we got here, this incredibly in-depth video from Bismuth highlights how the A-button challenge community managed to whittle down the number of A-presses, which players must use to make Mario jump or perform other actions, required to beat Super Mario 64.

The real (and sometimes controversial) science behind Apple's Extrapolations

JUSTINE GALMA MAR 17

- If you're diabetic, don't wait for your ametrivator to replace your needles (VICTORIA KONG MAR 18)
- How data centers at public pools can keep swimmers warm (JUSTINE GALMA MAR 18)

Facebook and Instagram's paid verification launches in the US

JAY PETERS MAR 17

- Donald Trump has started posting on YouTube again (ADI ROBERTSON MAR 17)
- Twitch CEO Emmett Shear is resigning (JAY PETERS MAR 18)

Page 8

Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>

And yes, it's five and a half hours long, but it's well worth the watch if you're interested in the game mechanics and intricate techniques that make this feat possible.



SECURITY Feds arrest alleged BreachForums owner linked to FBI hacks

EMMA ROTH MAR 15



EMMA ROTH MAR 15

Get ready, Washington. TikTok's planning to flood the nation's capital with "dozens" of influencers next week for a three-day protest of Biden's potential ban on the app, according to a report from *Politico*.

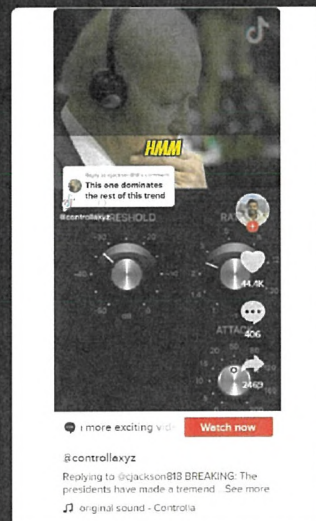
The Chinese-owned app is reportedly even paying the influencers for their journey to DC, although it's unclear who will be in attendance. TikTok spokesperson Jamal Brown has since confirmed the company's plans to *Politico*:

We look forward to welcoming our creators to our nation's capital, helping them make their voices heard, and continuing to drive meaningful impact in their lives and for their communities.

[TikTok's plan to stave off government intervention Flood D.C. with influencers](#)
 [POLITICO]

NILAY PATEL MAR 15

Presidents arguing about audio mixing is the best use of AI. As I mentioned on The Vergecast this week, these are my favorite AI jokes of all.



SECURITY Two hackers charged with last year's DEA portal breach

EMMA ROTH MAR 15

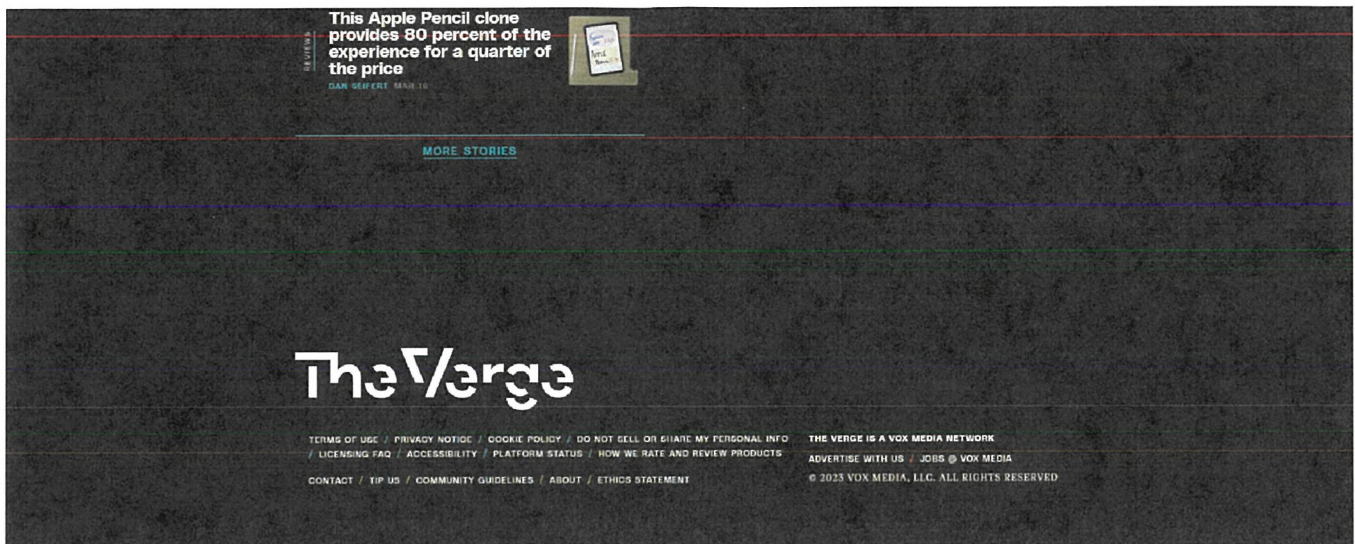


HEALTH If you're diabetic, don't wait for your smartwatch to replace your needles

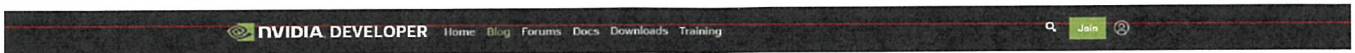
VICTORIA SONG MAR 15



Page 9
Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer - The Verge
<https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai>



Page 1
Facebook AI Researchers Achieve a 107x Speedup for Training Virtual Agents | NVIDIA Technical Blog
<https://developer.nvidia.com/blog/facebook-ai-researchers-achieve-a-107x-speedup-for-training-virtual-agents/>



Technical Blog

Subscribe

News

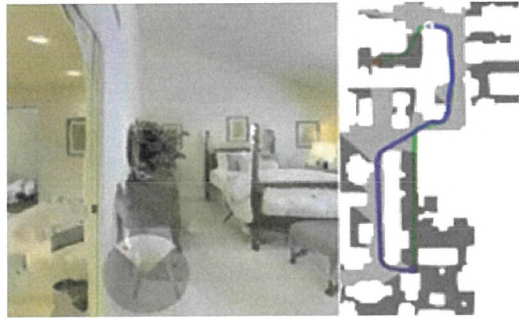
Jan 21, 2020

Facebook AI Researchers Achieve a 107x Speedup for Training Virtual Agents

By Heft Alarcon

Discuss (0) Like

Tags: featured, Machine Learning & Artificial Intelligence, News



Navigating a new indoor space without any prior knowledge or even a map is a challenging task for a human, let alone a robot.

To help develop **intelligent machines** that interact more effectively with complex 3D environments, Facebook researchers developed a GPU-accelerated **deep** reinforcement learning model that achieves near 100 percent success in navigating a variety of virtual environments without a pre-provided map.

To achieve this breakthrough, the team focused their work on developing an efficient approach to scaling RL models, which require a significant number of training samples, using multi-node distribution.

"A single parameter server and thousands of (typically CPU) workers may be fundamentally incompatible with the needs of modern computer vision and robotics communities," the researchers explained in their post. Near-perfect point-goal navigation from 2.5 billion frames of experience. "Unlike Gym or Atari, 3D simulators require GPU acceleration... The desired agents operate from high-dimensional inputs (pixels) and use deep networks, such as ResNet50, which strain the parameter server. Thus, existing distributed RL architectures do not scale and there is a need to develop a new distributed architecture."

Using **NVIDIA V100 GPUs**, with the **cuDNN**-accelerated **PyTorch** deep learning framework, and the **NVIDIA Collective Communications Library (NCCL)** in the backend, the researchers achieved a speedup of 107x over a serial implementation, by training their model on over 2.5 billion frames of experience.

"We leverage this scaling to train an agent for 2.5 billion steps of experience (the equivalent of 80 years of human experience) – over 6 months of GPU-time training in under 3 days of wall-clock time with 64 GPUs," the researchers stated in their paper, **Decentralized Distributed PPO: Solving PointGoal Navigation**, to be presented at ICLR 2020 in Ethiopia later this year.



In the paper, the team describes their decentralized method for scaling policy optimizations, aptly named **Decentralized Distributed Proximal Policy Optimization (DD-PPO)**.

In DD-PPO, each virtual agent alternates between collecting experience in a resource-intensive and GPU-accelerated simulated environment and optimizing the model.

Previous systems achieved a 92% success rate on these tasks. However, failing in the physical world can have serious ramifications, such as damaging a robot or its surroundings.

"DD-PPO-trained agents reach their goal 99.9 percent of the time. Perhaps even more impressively, they do so with near-maximal efficiency, choosing a path that comes within 3 percent (on average) of matching the shortest possible route from the starting point to the goal," the Facebook researchers stated in their newly published post on the **Facebook AI blog**.

"It is worth stressing how uncompromising this task is. There is no scope for mistakes of any kind — no wrong turn at a crossroads, no backtracking from a dead-end, no exploration or deviation of any kind from the most direct path."

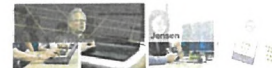
Topics

More topics

- + Computer Vision
- + Conversational AI
- + Cybersecurity
- + Data Center
- + Data Science
- + Generative AI
- + Networking
- + Recommenders
- + Rendering
- + Robotics
- + Simulation
- + Automatic Speech Recognition (ASR)
- + Automotive
- + CUDA
- + Digital Twin & Metaverse
- + Gaming
- + Healthcare & Life Sciences
- + HPC
- + News
- + NVIDIA Research
- + Omniverse
- + Technical Walkthrough

Don't Miss This: Defining Moment in AI
Keynote Premiere | March 21

Save this video



Related posts



Autonomous Machines
Transform the Future with Robotics at NVIDIA GTC



Robotics
Introducing NVIDIA Isaac Gym: End-to-End Reinforcement Learning for Robotics



Computer Vision / Video Analytics
AI at the Edge Challenge Spotlight: Sim-to-Real, an Effective Robot Navigation Framework



Computer Vision / Video Analytics
New Open Source GPU-Accelerated Atari Emulator for Reinforcement Learning Now Available



Robotics
Reinforcement Learning Algorithm Helps Train Thousands of Robots Simultaneously

Featured



Edge Computing
Upcoming Event: NVIDIA Jetson Edge AI Developer Days



Computer Vision / Video Analytics
Top Deep Learning Sessions at NVIDIA GTC 2023



Data Science
Maximizing Performance with Massively Parallel Hash Maps on GPUs



Simulation / Modeling / Design
Just Released: CUDA Toolkit 12.1

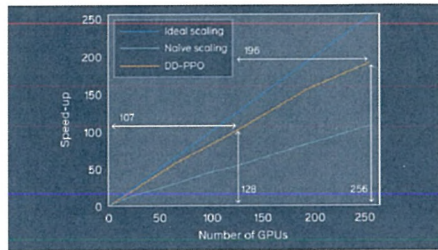


Conversational AI / NLP
Top Conversational AI Sessions at NVIDIA GTC 2023



Page 2

Facebook AI Researchers Achieve a 107x Speedup for Training Virtual Agents | NVIDIA Technical Blog
<https://developer.nvidia.com/blog/facebook-ai-researchers-achieve-a-107x-speedup-for-training-virtual-agents/>



DD-PPO demonstrates near-linear scaling as the number of GPUs increases from one to 250. Source: Facebook

The Facebook team trained and evaluated their model using the **AI Habitat** platform, an open source modular framework with a highly performant and stable simulator.

"Reaching billions of steps of experience not only sets the state of the art on the **Habitat Autonomous Navigation Challenge 2019** but also essentially solves the task," the researchers said. "It achieves a success rate of 99.9 percent and a score of 96.9 percent on the SPL efficiency metric. (SPL refers to success rate weighted by normalized inverse path length.)"

The team says they hope to build on DD-PPO's success by creating systems that accomplish point-goal navigation with only the camera input, and no compass or GPS data.

In addition to their in-depth explainer post, **Near-perfect point-goal navigation from 2.5 billion frames of experience**, the researchers have made the code publicly available on **GitHub**.

About the Authors



About Nefti Alarcon

Nefti Alarcon is a senior executive communications manager on NVIDIA's leadership team. He has years of media relations and communication experience, and has previously worked at Google, Mozilla, and CNN. He received his bachelor's degree in Journalism from George Washington University.

[View all posts by Nefti Alarcon »](#)

Comments

Start the discussion at forums.developer.nvidia.com



SIGN UP FOR NVIDIA DEVELOPER NEWS

[Subscribe](#)

Follow NVIDIA Developer



NVIDIA DEVELOPER

Copyright © 2023 NVIDIA Corporation

[Legal Information](#) | [Privacy Policy](#) | [Cookie Policy](#) | [Contact](#)

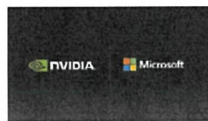


Accelerated Computing on Microsoft Azure

AI and High-Performance Computing for the Enterprise

Today's global challenges—including those related to our environment, economy, energy, and public health system—require modern, transformative solutions. Microsoft Azure and NVIDIA are empowering enterprises to push the boundaries of innovation by combining access to NVIDIA's full-stack computing platform with Microsoft's global scale, simplified infrastructure management, and flexibility to deploy from the cloud to the edge.

Explore Ways to Get More out of NVIDIA on Azure



NVIDIA and Microsoft Transform Cloud AI

First public cloud with complete NVIDIA AI solution stack combined with Azure's global scale for enterprise AI.

[Learn More >](#)



Access Technical Resources, On-Demand Webinars, and More

Microsoft Azure and NVIDIA deliver global access to accelerated computing on demand, simplified infrastructure management, and developer solutions that support the end-to-end lifecycle of building AI-powered applications.

[Learn More >](#)



NVIDIA and Microsoft Create Edge-to-Cloud Real-Time Streaming Video Analytics Solution

NVIDIA and Microsoft are partnering to enable real-time streaming and video analytics that extract powerful insights from thousands of cameras distributed over wide areas.

[Learn More >](#)



How to Launch NVIDIA RTX Virtual Workstations on Azure

An easy-to-follow guide takes you from creating your Microsoft Azure account to powering up an NVIDIA RTX Virtual Workstation (VWS) to access the most demanding design and engineering applications from the cloud.

[Watch Now >](#)

NVIDIA AI Enterprise on Microsoft Azure

GPU-accelerated instances on Microsoft Azure are certified and supported with **NVIDIA AI Enterprise**, a fully managed and secure, cloud-native suite of AI and data analytics software that streamlines each step of the AI workflow, from data processing and AI model training to simulation and large-scale deployment, reducing the time to move from pilot to production of AI solutions. It includes the broadly adopted software of the NVIDIA AI platform essential for developing predictive models to automate business processes and gain rapid business insights with applications such as conversational AI, recommender systems, **computer vision**, and more. NVIDIA AI Enterprise is certified on Azure with the following instances: NC-T4-v3, NC-v3, ND-A100-v4, NV-A10-v5.

NVIDIA GPU-Accelerated Virtual Machines on Microsoft Azure

Microsoft Azure and NVIDIA empower enterprises in the cloud to harness the combined power of NVIDIA accelerated computing and NVIDIA networking on demand to meet the diverse computational requirements of AI, machine learning, data analytics, graphics, virtual desktop, and high-performance computing (HPC) applications. From fractional GPUs and single GPUs to multiple GPUs across multiple nodes for distributed computing, access the right sized GPU acceleration for your workloads.

ND A100 v4 VM

Featuring eight NVIDIA A100 40GB Tensor Core GPUs, NVIDIA® NVLink® 3.0, and a dedicated NVIDIA Quantum 200 gigabits per second (Gb/s) InfiniBand connection per virtual machine (VM) for scale-out, multi-node, multi-GPU distributed computing.

Best suited for AI training, deep learning inference, machine learning, industrial HPC, and data analytics workloads.

[Learn More >](#)

NDm A100 v4 VM

Featuring eight NVIDIA A100 80GB Tensor Core GPUs with twice the GPU memory per VM compared to the ND A100 v4 VM series. Includes support for NVIDIA NVLink 3.0 and a NVIDIA Quantum 200 Gb/s InfiniBand connection per VM for scale-out, multi-node, multi-GPU distributed computing.

Best suited for recommender systems, distributed deep learning training, deep learning inference, machine learning, industrial HPC, and big data analytics.

[Learn More >](#)

NC A100 v4 VM

Offers the flexibility to select one, two, or four NVIDIA A100 80GB Tensor Core GPUs per VM to leverage the right-sized GPU acceleration for your workload. NVIDIA NVLink 3.0 is supported for GPU-to-GPU communication within the VM.

Best suited for single-node deep learning training, batch inference, interactive machine learning development and exploration, modeling, simulation, and data analytics.

[Learn More >](#)

NV A10 v5 VM

Offers the flexibility to provision partial GPU partitions to two full NVIDIA A10 Tensor Core GPUs per VM. Powered by Microsoft Azure GPU-partitioning capabilities built on top of NVIDIA RTX Virtual Workstation technology.

Best suited for graphics-intensive workloads, including virtual desktops, computer-aided design (CAD), rendering, simulation, AI inferencing, and data analytics.

[Learn More >](#)

Explore Azure Success Stories

Learn how companies like yours are creating value with NVIDIA on Azure



Microsoft Advancing AI-Powered Speech Using GPU Inference

Microsoft demonstrates how their voice search tools leverage NVIDIA inference on Azure to provide more accurate and human-sounding results to users 5X faster.

[Learn More >](#)



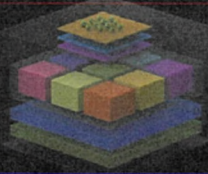
AI Helps Bing Search for Images Within Images

Search just got smarter, thanks to AI and NVIDIA GPUs on Azure. Microsoft's Bing now lets you search for images within images. You can even buy items you find there.

[Learn More >](#)

“Every industry has awoken to the potential of AI. We've worked with Microsoft to create a lightning-fast AI platform that is available to Microsoft Azure cloud users. With Microsoft's global reach, every company around the world can now tap the power of AI to transform their business.”

- Jensen Huang, CEO and Founder of NVIDIA




NGC

Get simple access to a broad range of performance-engineered containers for AI, HPC, and HPC visualization to run on Azure N-series machines from the NVIDIA NGC™ catalog. NGC containers include all necessary dependencies, such as NVIDIA CUDA® runtime, NVIDIA libraries, and an operating system, and they're tuned across the stack for optimal performance. To run NGC containers on Azure and take full advantage of NVIDIA GPUs, NVIDIA tested and validated containers are available on Azure Marketplace.

Azure Machine Learning Service

Use Azure Machine Learning service to accelerate the machine learning lifecycle with powerful NVIDIA GPUs. Use automated machine learning to identify suitable algorithms and tune hyperparameters faster. Improve productivity and reduce costs with autoscaling GPU clusters and built-in machine learning operations. Seamlessly deploy to the cloud and the edge with one click. Access all these capabilities from any Python environment using open-source frameworks such as PyTorch, TensorFlow, and scikit-learn. Azure Machine Learning service also integrates with NVIDIA Triton Inference Server and NVIDIA RAPIDS® to unlock even more performance gains.



GPU-Accelerated Virtualized Graphics

With NVIDIA RTX Virtual Workstations, creative and technical professionals can maximize their productivity from anywhere by accessing the most demanding professional design and engineering applications from the cloud. The latest NVads A10 v5 instances, powered by NVIDIA A10 Tensor Core GPUs, offer support for GPU virtualization and GPU-partitioning features, providing customers with access to scalable graphics and compute resources. Virtual workstations powered by NVIDIA GPUs are available directly from Microsoft Azure and the Azure Marketplace. For Remote Desktop Services (RDS) environments, knowledge workers and other professionals can leverage GPU-accelerated cloud computing on Windows multi-sessions with Azure Virtual Desktop.

Access the Power of Microsoft Azure and NVIDIA GPUs

Try the N-series today.

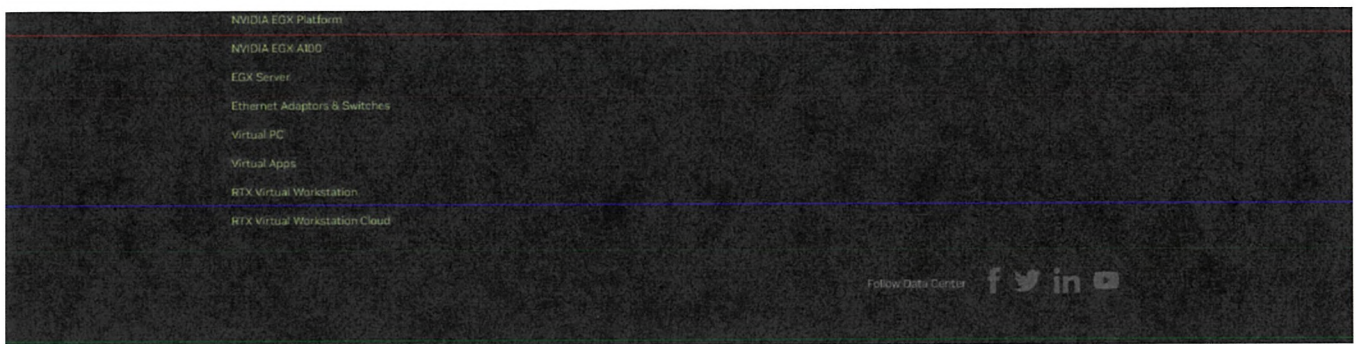
Get started with NGC.

[Learn More](#)

[Learn More](#)

Products	Technologies	Software	Resources
NVIDIA H100 CNX	NVIDIA Hopper Architecture	Overview	Data Center Blogs
NVIDIA H100	NVIDIA Ampere Architecture	NVIDIA AI Enterprise	GPU Apps Catalog
NVIDIA A100	Confidential Computing	Base Command	Data Center GPUs Product Literature
NVIDIA A2	NVLink-C2C	Bright Cluster Manager	Literature
NVIDIA A10	NVLink/NVSwitch	CUDA-X	DGX Product Literature
NVIDIA A16	Tensor Cores	Fleet Command	Virtual GPU Product Literature
NVIDIA A30	Multi-Instance GPU	Magnum IO	GPU Test Drive
NVIDIA A40	Index/ParaView Plugin	Networking	Where to Buy
NVIDIA L40	NVIDIA Morpheus AI framework	NGC Catalog	Qualified System Catalog
NVIDIA BlueField DPU		NVIDIA NGC	NVIDIA GRID Community Advisers
NVIDIA Converged Accelerators		Virtualization	Virtual GPU Forum
NVIDIA ConnectX SmartNIC			
NVIDIA V100			
NVIDIA HGX			
NVIDIA DGX H100			
NVIDIA DGX Systems			

Page 4
GPU Accelerated Computing on Microsoft Azure | NVIDIA
<https://www.nvidia.com/en-us/data-center/gpu-cloud-computing/microsoft-azure/>



United States

[Privacy Policy](#) | [Manage My Privacy](#) | [Legal](#) | [Accessibility](#) | [Corporate Policies](#) | [Product Security](#) | [Contact](#)
Copyright © 2023 NVIDIA Corporation